Cluster-driven evolution and modularity uncover paths of cholera emergence

Mario López-Pérez a,b,c, Trudy-Ann Grant a,b,Jose M. Haro-Moreno c,d, Asier Zaragoza-Solas c, Salvador Almagro-Moreno a,b.*

aBurnett School of Biomedical Sciences, College of Medicine, University of Central Florida, Orlando, Florida, USA; bNational Center for Integrated Coastal Research, University of Central Florida, Orlando, Florida, USA; cEvolutionary Genomics Group, División de Microbiología, Universidad Miguel Hernández, Alicante, Spain.

*Address correspondence to: Salvador Almagro-Moreno, samoren@ucf.edu

# Contributed equally

Running title: Cluster-driven evolution of Vibrio cholerae

Keywords:
Pathogen emergence / Vibrio cholerae / aquatic pathogen / virulence evolution / population genomics / cluster divergence
ABSTRACT

Pathogen emergence remains one of the most pressing public health concerns of our times. Here, using the agent of cholera, the *Vibrio cholerae* pandemic cholera group (PCG) as a model system, we investigate the evolutionary dynamics that lead to the emergence of human pathogens from environmental populations. Genomic comparison of over 1,100 *V. cholerae* genomes including novel isolates from this study, reveal a generalized cluster-driven phylogeny and evolution of the species. Emergence of PCG is largely based on modular acquisition of mobile genetic elements including small gene clusters and allelic variations, with distinct phylogenomic clusters acting as differential reservoirs of virulence factors. Surprisingly, PCG encodes few unique genes, which are mostly encoded within the super-integron, however, many genes within the group exhibit an aberrant degree of positive selection. Our analyses provide a compelling scenario for the emergence of pandemic clones in *V. cholerae* establishing a blueprint for other bacterial pathogens.
INTRODUCTION

The emergence of human pathogens is one of the most pressing and persistent public health concerns of modern times, with new outbreaks of varying magnitude occurring regularly around the world\textsuperscript{1–4}. Pathogen emergence is a complex and multifactorial phenomenon that culminates with a microorganism acquiring the ability to colonize and harm the human host, cause outbreaks and persist in the environment between them\textsuperscript{5}. To date, the specific factors that allow certain strains within a species to emerge as pathogens remain mostly enigmatic\textsuperscript{4}. Elucidating how the interplay between the numerous molecular and environmental drivers lead to this complex phenomenon is critical to develop surveillance platforms to predict emergent events and potential sources of outbreaks\textsuperscript{4}.

The Vibrionaceae is a highly diverse family of aquatic bacteria that encompass several species that represent distinct paradigms of how pathogens can emerge from environmental populations\textsuperscript{6,7}. \textit{Vibrio cholerae} is the most widely studied member of the Vibrionaceae and has been a model organism for countless molecular, ecological and epidemiological studies. \textit{V. cholerae} is the causative agent of the severe diarrheal disease cholera, a scourge that affects millions of people each year and is responsible for over 100,000 deaths annually\textsuperscript{8}. The disease typically affects nations with limited access to clean water and poor sanitation and can be lethal if not properly treated\textsuperscript{8,9}. Cholera has existed in the Indian subcontinent for centuries, however, the disease started to spread worldwide after 1817 causing seven pandemics\textsuperscript{8–10}. The first six pandemics are thought to have been caused by the Classical biotype \textit{V. cholerae} O1, a strain that is now considered virtually extinct\textsuperscript{8–11}. The seventh and current pandemic is caused by the El Tor biotype of \textit{V. cholerae} O1 and started in Indonesia in 1961\textsuperscript{12,13}. The recent cholera epidemics have increased in duration, intensity, and frequency warranting urgent attention to this infectious disease\textsuperscript{8,9,14}. Furthermore, extended periods of warm weather, driven by climate change, have provided suitable conditions for the proliferation of pathogenic \textit{Vibrio} spp.\textsuperscript{14,15}. For instance, a recent multidecadal study demonstrated a
steady increase in the abundance of pathogenic Vibrios, including *V. cholerae*, over the past half-century\textsuperscript{16}. Overall, these concerning patterns only exacerbate the problem of the emergence and reemergence of pathogenic Vibrios, the spread of virulence genes and their proliferation\textsuperscript{17}.

Interestingly, only a limited number of *V. cholerae* strains can cause pandemic cholera\textsuperscript{18–20}. Specifically, only those confined to a phylogenetic group, the pandemic cholera group (PCG), can cause the disease\textsuperscript{19,21,22}. Even though the mechanisms of cholera pathogenesis in humans have been extensively studied over the last several decades, the evolutionary events and molecular constraints that explain why only members from PCG can cause pandemic cholera remain unknown\textsuperscript{23–28}. Some of the major general events that are essential for the emergence of pandemic *V. cholerae* from environmental populations include the acquisition of the lipopolysaccharide O1 antigen cluster and the mobile genetic elements (MGEs) such as CTXΦ phage, *Vibrio* pathogenicity Island-1 (VPI-1) and *Vibrio* pathogenicity Island-2 (VPI-2)\textsuperscript{11}. Among several other differences with Classical, El Tor strains have also acquired two unique pathogenicity islands (PAIs): *Vibrio* Seventh Pandemic (VSP) 1 and 2. CTXΦ encodes the genes for the cholera toxin, which is responsible for the severity of the profuse diarrhea associated with cholera\textsuperscript{29}. The toxin coregulated pilus (TCP), an essential colonization factor, is encoded within VPI-1\textsuperscript{30,31}. This island also encodes several major virulence regulators such as ToxT and TcpP\textsuperscript{31,32}. VPI-2, among other, encodes the genes for sialic acid utilization, which confers a competitive advantage in the gut\textsuperscript{33}. VSP-I encodes a regulator that plays a role in intestinal colonization and chemotaxis\textsuperscript{34}. VSP-II is associated with environmental survival and fitness of the bacterium\textsuperscript{35–37}. It was recently found that VSP-II and VPI-2 encode systems that prevent the uptake of foreign genetic material\textsuperscript{38}. These MGEs are always found in the PCG, however, they are not exclusive to them as they are also encoded in some environmental strains of *V. cholerae*\textsuperscript{39–41}. However, their abundance, specific distribution and dynamics within the *V. cholerae* species remains poorly understood.
Horizontal acquisition of key virulence genes is a critical step in the emergence of toxigenic clones of *V. cholerae*, however, it is not sufficient to explain why the ability to cause cholera is limited to clones belonging to the PCG. This phylogenetically confined distribution together with the presence of CTXΦ and the PAIs in some environmental strains strongly indicate that other barriers and requirements exist that stringently limit their emergence. To begin determining these barriers and the evolutionary origins of pandemic clones of *V. cholerae*, we recently investigated the genomic prerequisites that must be present in a population before a pathogenic clone can emerge from an environmental gene pool. We determined that the environmental ancestor of the pandemic group had a particular genomic background containing a specific set of alleles of core genes that conferred preadaptations to virulence and enhanced its potential to give rise to pandemic disease, what we term virulence adaptive polymorphisms (VAPs). Interestingly, our data shows that VAPs are present in environmental populations and confer virulence-associated traits to strains encoding them. Overall, our results indicate that there are core genomic factors beyond the acquisition of MGEs that are critical for the emergence of pandemic *V. cholerae*. To date, the unique core genes and allelic variations in the genomic background of pandemic *V. cholerae* that explain their narrow distribution remain enigmatic.

In this study, we investigated the evolutionary paths that led with the emergence of pandemic *V. cholerae* from environmental populations. To do this, first, we isolated environmental *V. cholerae* strains from the Indian River Lagoon (IRL), a highly diverse and expansive estuary located in Eastern Florida. Genome comparisons of these strains against *V. cholerae* PCG unveil a phylogeny in which strains form cluster-like groups following a similar pattern as what we recently found with *Vibrio vulnificus*. We extended our analyses to over 1,100 publicly available genomes and found that *V. cholerae* strains are divided into four major clusters and numerous other minor clades. We found that PCG belong to one large cluster (C24) and is located within one lineage shared with among other environmental strains, the pandemic cholera lineage (PCL).
This topology suggests a scenario where clusters may differ in their virulence gene content. In order to test this, first, we first investigated the evolutionary history, environmental reservoirs and abundance of the O1 capsule and MGEs associated with cholera emergence. We found widespread modularity of these elements, with distinct phylogenomic clusters acting as differential reservoirs of virulence factors. However, certain allelic variations and scarcely distributed islands act as bottlenecks drastically narrowing the distribution of PCG-like gene clusters. Furthermore, PCG encodes very few unique genes, which are mostly encoded within the super-integron, however, many genes within the group exhibit an aberrant degree of positive selection, which may further explain the confined distribution of these isolates. Our analyses provide a compelling scenario for the emergence of pandemic clones in V. cholerae establishing a blueprint for other bacterial pathogens.

RESULTS AND DISCUSSION

Isolation of Vibrio cholerae in the Indian River Lagoon. V. cholerae is a natural inhabitant of aquatic ecosystems where it can be found as free-living or attached to abiotic or biotic surfaces. We recently established two sampling sites along the eastern coast of Florida where we isolated pathogenic strains of Vibrio vulnificus. We sampled those two sites, Feller’s House (FH) and Shepards Park (SP), to investigate the potential presence of V. cholerae strains (Fig. 1A). Using the isolation, enrichment and selection methods described below, we identified a total of 178 potential V. cholerae isolates from the different fractions at both locations during our sampling events (42 FH and 136 SP). In order to confirm those isolates, we performed a PCR screening for the species-specific gene ompW. Interestingly, only 65 isolates were confirmed as bona-fide V. cholerae all of them isolated from SP (Fig. 1A). V. cholerae isolates from SP were collected from all the fractions that we examined (sediment, cyanobacteria, particle-associated and water) except for oysters. Subsequently, we performed PCR-based fingerprinting of the strains to reduce redundancies and minimize clonality. We used a combination of PCR-based
fingerprinting analyses, VCR-PCR and ERIC-PCR. Based on fingerprinting patterns and cluster analysis performed using BIONUMERICS v8.0 we generated a dendrogram to group these observed patterns. Using this resulting dataset, we selected 35 isolates with unique fingerprinting patterns for whole genome sequencing (Fig. 1B and S1).

Phylogenomic and clustering of V. cholerae strains. The 35 V. cholerae strains that we sequenced were isolated from a diverse set of fractions: 6 from sediment, 10 from cyanobacteria, 9 from the free-living water fraction and 10 from the particle-associated one (Table S1). We examined the relationship of these strains with V. cholerae PCG by analyzing the average nucleotide identity (ANI) in a pairwise comparison against V. cholerae N16961, an O1 strain from the El Tor biotype (Fig. S2). The minimum values obtained were ca. 97.5, always within the established cut-off for strains of the same species. However, up to nine environmental strains showed a greater divergence with the PCG (ANI > 98.2%) than with the rest of the strains (Fig. S2). Interestingly, phylogenomic tree analysis based on whole-genome single nucleotide polymorphism (SNP) (230,824 SNPs identified in 3,19 Mb of core genome) show that the majority of the IRL strains form two major clusters one of them including the toxigenic strains of V. cholerae (Fig. 1B). This evolutionary relationship shares similarity with that of V. vulnificus, in which the pathogenic strains capable of causing septicemia in humans belong to one clearly defined cluster. This prompted us to characterize the V. cholerae species to investigate whether the evolution of the pandemic strains follow a similar cluster-based pattern as this would shed important light in the emergence and evolution of the pathogen.

First, we collected the V. cholerae genomes available in public databases (based on NCBI classification accessed in July 2020; see Materials and Methods) together with our environmental isolates. Naturally, there is a significant bias in the databases towards V. cholerae O1 and O139 strains, which are mostly clonal. Furthermore, it is likely that several V. cholerae environmental strains that were uploaded in the databases are also
from clonal origin due to enrichment during the early steps of isolation. Therefore, in order to obtain an unbiased phylogeny, we dereplicated all these genomes to an identity of ANI>99%. From a total of 1,127 initial genomes, we obtained 227 non-clonal reference \textit{V. cholerae} clades (Table S2). To maximize the consistency of our analyses, we built both a SNP-based phylogeny and genome clustering using ANI to determine the groups that make up the \textit{V. cholerae} species. Clustering using a threshold of 98% of ANI shows that the species encompasses 34 clusters that include \textit{V. cholerae} strains (Table S3). The application of an ANI>97% results in a single cluster indicating that boundary of intra-population sequence diversity in \textit{V. cholerae} species is 97%. Next, we used a total of 458,798 SNPs from a core genome size of 2.46 Mb to build the species phylogeny using non-clonal reference clades and IRL strains (Fig. 1C). The species phylogeny topology based on SNPs or ANI clustering are virtually identical, with some minor exceptions that might be due to certain strains having a convergent history. Interestingly, both approximations reveal that four major clusters account for 69.2% of the non-clonal clades (157 out of a total of 227) including 31 of the 35 IRL isolates from this study (Table S3). All IRL isolates are grouped together with previously isolated strains except for strains IRLE0050 and IRLE0074 which formed a separate cluster (Table S3). Cluster 24 (C24) is the largest of the four main clusters containing 68 non-clonal clades including PCG. On the other hand, most IRL strains belong to C5 (22 of the 35) (Fig. 1C). We did not find a pattern that associate the phylogeny of the IRL isolates and their fraction of isolation. The C24 phylogeny also reveals that PCG groups with nine other non-clonal clades, including one from this study IRLE0178, into a separate branch in the phylogenomic tree, what we have termed the Pandemic Cholera Lineage (PCL) (Fig. 1C). This provides us with a novel hierarchical approach based on three levels to decipher the evolutionary history of pandemic strains of \textit{V. cholerae}: (from highest to lowest) a) the C24 within all \textit{V. cholerae} strains, b) the PCL within C24 non-clonal clades and c) the clonal \textit{V. cholerae} pandemic strains (PCG) within the PCL lineage.
Lipopolysaccharide (LPS) cluster. The LPS plays numerous roles in the pathogenesis of pandemic *V. cholerae* (e.g., endotoxic activities, immunological responses, etc.) and is critical to determine the antigenic properties and classification of the bacterium. Specifically, given that only O1 strains, together with the extinct O139, can cause pandemic cholera we examined its LPS cluster to determine its distribution and evolutionary history within the *V. cholerae* species (Fig. 2). The O1 gene cluster is 29.7Kb long and encompasses approximately 26 ORFs with potential or demonstrated function (some are truncated preventing an exact number). The *rfa* gene cluster (LPS core synthesis) is encoded at the 5' end of the island (Fig. 2A blue arrows). The *rfa* cluster is widely distributed across all clades, with the notable exception of RfaF (ADP-heptose--LPS heptosyltransferase 2, VC_0235) and RfaL (O-antigen ligase, VC_0237) present in ca. 35 non-clonal clades. Nonetheless, the genes in the *rfa* cluster exhibit the highest divergence within the LPS cluster (median identity 86-88%) and accumulate numerous synonymous polymorphisms (Fig. 2A). Indicating that, like *V. vulnificus*, they appear to be a preferred site to act as hot spots for the recombination of this island due to the conservation and synteny of these genes. The distribution of the central region of the LPS cluster is much more restricted and is only present in 17 *V. cholerae* non-clonal clades, however, those that encode it share 100% identity with the region of PCG. This region contains the *rfb* cluster which is associated with antigen-O synthesis (Fig. 2A, green arrows). Interestingly, only 5 out the 17 clades that encode this region belong to the PCL, indicating that there is no direct correlation between the phylogeny and its presence in PCG. The LPS cluster of PCG encodes four transposases, which likely increase the variability within the island (Fig. 2A and 2B). These IS elements are widely distributed among *V. cholerae* strains providing insight into how small gene clusters can be introduced in the island. For instance, even though strains VC22 (C24) and 2479-86 (C27) encode the same version of the PCG LPS cluster (identity >99%), they do not encode two of these transposases (Fig. 2B). Interestingly, these 2 transposases are inserted within *rfbT* in PCG, leaving a truncated version of the gene (Fig. 2B and 2C).
rfbT encodes a methyltransferase that is responsible for seroconversion between V. cholerae O1 Ogawa and Inaba serotypes, suggesting a major role for these IS sequences in the emergence of PCG. Furthermore, several strains from C23 encode the same version of the hypervariable region as VC22 and 2479-86 (identity 99%) but the rfa cluster is different (Fig. 2B). The first three genes of the island encode different alleles of rfaF, rfaG and rfaL, and the identity of the remaining three is 95% (Fig. 2B). The rfa cluster pattern of these strains is similar to that of strain A325 (C30) (Fig. 2B). In addition, in this strain the two transposases that truncate the rfbT gene in the PCG are inverted maintaining an intact version of the gene (Fig. 2C). Overall, we found a remarkable degree of conservation of the LPS cluster, which is present in almost twenty clades encoding a PCG-like version. This is particularly striking as capsular hypervariability is a universal bacterial characteristic, suggesting some strong evolutionary pressure maintaining this version of the island widespread in the population.

It appears that the island is a) exchanged by homologous recombination leading to the whole gene cluster being conserved and b) the presence of IS elements increases the variability by truncating genes such as rfbT.

Abundance and distribution of virulence-associated mobile genetic elements in V. cholerae species. There are several horizontally acquired MGEs that are directly associated with the emergence of pandemic clones of V. cholerae. After we established the phylogeny and clustering of the V. cholerae species, we investigated the distribution, evolutionary dynamics and composition of the different virulence-associated MGEs.

1. CTXΦ phage. The genes for cholera toxin (CT) are encoded within the CTXΦ phage. The production of CT in the intestine is responsible for the severity of the profuse diarrhea associated with cholera. CT leads to the secretion of electrolytes and nutrients such as fatty acids into the lumen of the small intestine. Due to the lysogenic nature of CTXΦ, the insertion and deletion dynamics of this phage enables
recombination of genes, which leads to diversity within the pandemic strains ultimately serving as an opportunity to increase the pathogenic potential of pandemic strains\textsuperscript{29,55,56}. Inquiry into the distribution and composition of this element within \textit{V. cholerae} shows that there is no environmental strain that recapitulates all the phage genes of PCG. At most, five clades and two clusters encode 10 of the 14 genes that make up the CTX\(\Phi\) phage of \textit{V. cholerae} N16961: three from C24 (represented by AAS91, PivertUAT4Aug and V060002) and two from C17 (3541-04 and 3528-08) (\textbf{Fig. 3A}). Four of these strains encode the two subunits of the cholera toxin (CtxAB). However, genes coding for the duplicated transcriptional regulator RstR (VC_1455 and VC_1464) are only found in one strain: AAS91 (\textbf{Fig. 3A}, brown shade). This strain belongs to the PCL and encodes the complete version of CTX\(\Phi\) except for VC_1460. This gene encodes a minor coat protein pIII related to host interaction as part of the infection mechanism (\textbf{Fig. 3B})\textsuperscript{57}. Instead, a transposase has been inserted in this locus in strain AAS91, followed by a APH(3')-II family aminoglycoside O-phosphotransferase (\textbf{Fig. 3B}). In addition, AAS91 has a third copy of the triplet (RstA-RstB-RstR) instead of the two copies present in PCG strains (\textbf{Fig. 3B}). The second most complete version of CTX\(\Phi\) in \textit{V. cholerae} belongs to strains V060002 from PCL and 3528-08 (C17). Both strains have lost the gene coding for RstC and encode a different allele of the transcriptional regulators RstR (\textbf{Fig. 3B}).

Genomic comparison of \textit{V. cholerae} strains reveal that the phage variation dynamics are of the "additive" kind, where small gene cassettes are integrated into the 3' part of the gene coding for RtxA\textsuperscript{58}. All the CTX\(\Phi\) variants have conserved intergenic regions that separate and mark the insertion of these gene cassettes (\textit{att} sites, marked as black arrows). These types of insertions have been reported previously in \textit{V. cholerae} O139\textsuperscript{59}. Interestingly, some strains such as 3528-08 (C17) can have up to five insertions (\textbf{Fig. 3C}). The first two of the five cassettes are identical, and each cassette has a copy of the two subunits of CT. The other three have lost the toxin genes and those coding for RstB, RstR and RstA, and acquired a replication endonuclease (\textbf{Fig. 3C}).
presence/absence of these cassettes is not correlated with the phylogeny and, when present, they have a high identity indicating a fast turnover of these elements (Fig. 3B). Furthermore, these cassettes can be integrated into other genome locations, as in the case of the strain PivertUAT4Aug (Fig. 3B). In addition, using the sequence of att sites as reference we also found a collection of cassettes in the same order (including the one containing the two subunits of CT) in the close relative species Vibrio mimicus, suggesting another potential donor of CTXΦ to non-toxigenic V. cholerae (Fig S2).

2. Vibrio Pathogenicity Island-1 (VPI-1). The toxin coregulated pilus (TCP), an essential colonization factor, is encoded within VPI-1. TCP mediates microcolony formation, which is crucial for intestinal colonization, and acts as the receptor of CTXΦ. VPI-1 also encodes two critical virulence regulators: ToxT and TcpP. VPI-1 can be transferred, via generalized transduction, between strains of V. cholerae and can also form circular intermediates. The variable region of the island is located between the CDS VC_0821 (hypothetical protein) and VC_0846 (integrase) (Fig. 4A). The most conserved version of this region can be found in nine clades, four of them within C24 (2 from PCL) and five from C17, including strain IRLE0081 isolated in this study (Fig. 1C). The average identity of the proteins was highly conserved compared to PCG (ca. >97%), however, we found some notable exceptions (Fig. 4A). First, among the proteins that are part of the TCP operon, we found high divergence rates in the master regulator ToxT (VC_0838) (Fig. 4C) and the toxin-coregulated pilus major pilin, TcpA (VC_0828) (Fig. 4D) with a median protein identity ca. 81% (Fig. 4). In addition, we found that the PCG allele of the colonization factor TcpF (VC_0837) (Fig. 4E) was only present in three strains (one PCL and two C17), in the rest, despite the conservation of synteny and similarity of the whole TCP operon, the identity of this protein was less than 40%, likely encoding a different allele of this poorly understood secreted protein. Finally, our genomic comparisons reveal that the gene coding for ToxR-activated gene A, TagA (VC_0820) was absent from the C17 genomes, while being retained in three of
the four C24 genomes (Fig. 4B). Overall, despite widespread conservation in the synteny and size of the clusters of genes that comprise VPI-1 in the *V. cholerae* species, the existence of unique major allelic variations in a limited number of genes act as a stringent bottleneck in the emergence of PCG.

3. *Vibrio Pathogenicity Island-2 (VPI-2).* VPI-2 is a large PAI that encodes the genes for sialic acid utilization. The capacity to utilize sialic acid as a carbon and energy source confers *V. cholerae* a competitive advantage in the mucus-rich environment of the gut, where sialic acid availability is extensive. Sialic acid catabolism also mediates a chemotactic response towards mucin and several environmental reservoirs of *V. cholerae*. VPI-2 is excisable and form a circular intermediate. Interestingly, during the excision event, there is cross-talk between VPI-2 and VPI-1. Unlike other MGEs, the differential abundance of the gene clusters within VPI-2 suggests that the island is made up of different modules. Also, all but one of the IS elements of the island are widespread among several clades in diverse locations, which favours the transfer of the different modules. The canonical VPI-2 from O1 strains encodes three major modules, from 5' to 3': a) a restriction modification system (RM), b) the Nan-Nag cluster, and c) a Mu-phage like region (Fig. 4F). At the 3' end of VPI-2 there is also a small group of genes encoding hypothetical proteins. The Nan-Nag cluster is present in 88 clades, whereas the distribution of the RM system is more limited, being encoded by only 7 clades (Fig. 4F). The Mu-phage like module is present in nine clades, however, there is no relationship between specific modules and species phylogeny. Like VPI-1, no clade other than PCG encode a PCG-like version of VPI-2. Furthermore, strains that contain the three individual modules sometimes encoded them in different locations within the genome. For instance, strain PivertUAT4Aug (C24), the only one with the three major modules, encode the first two modules (Nan-Nag and RM) within a version of VPI-2, whereas the Mu-phage region is encoded in a different location in the genome (Fig. 4G). Besides several genes missing, PivertUAT4Aug encodes a different allele of *hsdS*
Another instance of this modular distribution is exemplified by strain N2824 (C16), which possesses the first two modules but not the Mu-phage. Furthermore, the strain has lost the genes encoding the HsdS and HsdM proteins from the RM cluster involved in DNA sequence recognition and methylation, respectively (Fig. 4G). Strain 3541-04 (C17) encodes a different cluster of genes at the insertion site of the Nan-Nag cluster and RM system, with the Nan-Nag cluster being in another chromosomal region (Fig. 4G). Although this strain is the only one that contains a small cluster of four genes just like PCG and in the same position (from VC_1801 to VC_1804). Besides the gene mentioned above that codes for HsdS (VC_1768), there are three others that code for hypothetical proteins (VC_1787, VC_1788 and VC_1791) that are only present in the PCG (highlighted in Fig. 4F). Genomic comparisons with PivertUAT4Aug (C24) show that, similar to rfbT in the LPS gene cluster, VC_1788 and VC_1791 are the two parts of a gene coding for the tape measure protein of phage Mu that have been truncated by the insertion of two transposases in the PCG, annotated as VC_1789 and VC_1790 (Fig. 4G). Overall, VPI-2 in PCG must be understood based on its unique modular arrangement not by the individual units that constitute it. Its four modules are associated with IS elements that are widely distributed among V. cholerae environmental strains and scattered throughout the genome. This leads to the convergence of all modules in the same genome being extremely low which contributes to the exceptionality of PCG.

4. Vibrio Seventh Pandemic Island I (VSP-I). The role of VSP-I in V. cholerae pathogenesis is not as clearly defined as other MGEs. VSP-I has been suggested to play an environmental role related to chemotaxis and encodes a regulator involved in intestinal colonization (VC_0177). We found homologs of the eleven genes that comprise this island in ten V. cholerae clades (Fig. 5A). Although four of them are found in strains isolated from this study, IRLE0062, IRLE0049, IRLE0042 and IRLE0077, that belonged to the same clade. The greatest sequence divergence corresponded to an
integrase (VC_0183) and an XRE family transcriptional regulator (VC_0176) (Fig. 5A).

Nonetheless, only two strains encode a PCG-like VSP-I, V060002, which is phylogenetically closer to PCG (C24), and strain VcCHNf4 from C5. V060002 has two identical copies of the island in the genome that are in close proximity of each other in the genome with the LPS cluster located between them (Fig. 5B). Another strain belonging to C24 (2000), has the first gene at the 5’ end replaced with a cluster of four genes (Fig. 5B). The landscape of evolutionary possibilities for VSP-I is much more limited than for the other MGEs due to its small size, nonetheless, and despite of this, it is the rarest of all the islands. Suggesting this is a critical bottleneck in the emergence of the current El Tor strains.

5. Vibrio Seventh Pandemic Island II (VSP-II). The presence of VSP-II in clinical and environmental strains might be associated with environmental survival and fitness of the bacterium.35–37 A group of V. cholerae strains that caused an outbreak in Florida associated with oyster consumption, encoded a novel bacteriocin and a pyocin protein in the VSP-II element, which are typically associated with bacterial competition.37 Our analyses indicate that, similar to VPI-2, VSP-II is highly modular with several clusters of genes that have a tRNA-Met insertion point. The direct repetition of the end of the tRNA (highlighted as purple triangles) marks the insertion points of each module, which in most cases match the presence of IS elements (Fig. 5C and 5D). Overall, this island serves as an example of an "additive" island with three small modules. Module 1 comprises genes VC_0490 to VC_0498, five of which are of viral origin but lack essential proteins for packaging, indicating it is a defective prophage. Module 2 encompasses from VC_0502 to VC_0510, however, the lack of possible annotation of its genes precludes from assigning a specific function to this module. Nonetheless, VC_0502 and VC_0503 from this module are only present in strains from C24 (Fig. 5C). Module 3 is comprised of four genes (VC_0512 to VC_0515): VC_0512 and VC_0514 encode methyl-accepting chemotaxis proteins, VC_0513 a transcriptional regulator and VC_0515 a predicted
signal transduction protein. These four genes are only present in C24 strains and according to their annotation could be associated with environmental adaptations (Fig. 5D). Genomic comparison with strain MZO-3 (C24) reveal the presence of a novel CDS in the PCG between VC_0512 and VC_513 that codes for a TnsD family Tn7-like transposition protein (Fig. 5D). Overall, there are no identical versions of VSP-II outside of PCG. The closest versions are encoded by strains N2723, MZO-3, and V060002 (C24), which are the only ones that possess a similar PCG-like modularity and synteny of VSP-II (Fig. 5D). Relative to PCG, the changes in VSP-II from strain MZO-3 involve only IS elements, the loss of a transposase (VC_0501) in Module 1 and the gain of another in Module 3 (Fig. 5D). Whereas strain N2723 has lost genes coding for two hypothetical proteins (VC_0496 and VC_0509) and a transcriptional regulator (VC_0497) (Fig. 5B). In the rest of V. cholerae strains, VSP-II only maintains some of these modules or clusters of genes, with the addition of other types of clusters different from that of the PCG being common.

Overall, our analyses of the virulence-associated MGEs indicate that a) strains from V. cholerae PCG have acquired the gene clusters encoding the major virulence genes in a modular fashion, b) those MGEs are more widespread than previously thought within the V. cholerae species, and c) a select set of genes dispersed in the phylogeny of the species have a very limited distribution and d) unique allelic variations in PCG act as stringent bottlenecks for emergence.

Pangenome analysis. Besides the modular acquisition of MGEs, are there other genes associated with the emergence of PCG? In order to start addressing this question, we investigated the unique genes encoded by PCG and those that the group inherited from its closer ancestors. We considered the three major hierarchical levels in which the pandemic strains of V. cholerae are included: cluster (C24), pandemic cholera lineage (PCL), and the pandemic cholera group (PCG) (Fig. 1C). We analyzed the sequences
defining each of these groups at the pangenome scale, using the remaining set of clades as outgroups to remove the background of other environmental *V. cholerae* sequences. The Venn diagram in **Figure 6A** shows the total unique and shared gene family clusters among the pangenomes of the different groups analyzed. The total number of shared clusters of the *V. cholerae* species is 3,278 (core genome). The PCG encodes 29 unique gene families, 62% of which (18/29) are located in the integron of chromosome II (**Fig. 6A**). Most of the genes could not be assigned a function or annotation, nonetheless, among these unique gene families we found a transposase, a pullulanase, a lysozyme inhibitor LprI family protein and several acetyltransferases (**Table S4**). Among the remaining specific clusters that are encoded within the integron, most of them are localized in MGEs. PCG inherited ten genes from the PCL. Half of those are also concentrated in the integron and three are present in PAIs: the two homologs of the transcriptional repressor RstR of *CTXΦ* (VC_1455 and VC_1464) and the ToxR-activated gene A from VPI-1 (VC_0820) (**Table S4**). Finally, PCG inherited 17 unique gene families from C24 with more than 80% of them being encoded within the integron and the MGEs (**Fig. 6A**). Among these 17 genes we found two small clusters located in VSP-II (VC_0502-VC_0503 and VC_0512-VC_0515) mentioned above in the island description, as well as a cluster of genes inserted downstream of the 5' of *CTXΦ* (**Table S4**). This cluster encodes a helicase, a transcriptional regulator, and a phage replication protein, along with a Xer recombination activation factor involved in fragment integration into the chromosome. Overall, our pangenome analyses reveal that, besides the acquisition and modularity of MGEs, the integron plays a most critical role in the emergence and evolution of the PCG significantly greater than the absence or presence of other genes in the core genome.

**Evolution of the Pandemic Cholera Lineage.** Next, in order to understand at fine resolution the evolution of PCG from the root of the PCL, we paired a) the history of MGE acquisition, b) module exchange and c) gene gain and loss of the entire lineage.
Our analyses indicate that the capsule is the only element that is not modular in the PCL, which contains 10 clades including the PCG, and appears to have evolved early in the branching of PCL. It is present in 6 of the 9 non-PCG strains and is directly associated with branching event 3 (Fig. 6B). VPI-1 and CTXΦ also appear to have been acquired during branching event 3 and is present in the 2 clades that branch with PCG. On the other hand, VSP-I and VSP-II are only found in 1 strain each, none of them with a direct branching pattern related to PCG, which suggests independent acquisition events of the islands by PCG (Fig. 6B). Only the closest relative to PCG encodes VPI-2 suggesting a late acquisition of the island, likely during branching event 5. Overall, our results demonstrate the mosaic-like relationship between mobile element acquisition and the emergence of pandemic cholera strains with some (e.g. capsule) appearing early in the phylogeny and others (e.g. VPI-2) being late acquisitions.

Given the importance of branching event 3 in the evolution of PCG, we analysed the gene gain and loss of the different branches during the evolutionary process that led to the PCG group excluding those genes related to PAIs. Our analyses highlight the acquisition of small gene clusters by PCG during the evolution of its lineage after the split in node 3 (Table S5). Several of these genes colocalize consecutively in the integron suggesting a common evolutionary history. On chromosome I, we found a cluster of genes related to sugar transfer, such as a glycosyltransferase, possibly involved in cell wall decoration (Table S5). Two clusters were associated with transporters such as ion channels and type IV toxin/antitoxin systems. On chromosome II, apart from the integron, we found gene clusters associated with a transposase related to regulatory proteins and one linked to a peptidase. On the other hand, the genes lost in relation to the other PCL branches by PCG are related to membrane proteins such as one of the OmpA family, as well as several secretion systems and transcriptional regulators (Table S5). Interestingly, we found that PCG had lost a chitinase, which is a critical adaptation to the environment, the reason behind this remains to be elucidated.
Genetic variations associated with pathogen emergence (GVPEs). From our analyses it appears that differences in overall gene content minimally account for the confined distribution of pandemic cholera clones other than genes within the well-known MGEs and virulence-associated clusters. In light of this, we investigated whether allelic variations of core genes could help explain this confined distribution. We examined the core genome microdiversity of PCG by estimating the ratio of nonsynonymous (dN) to synonymous (dS) substitutions of each gene of the cholera group against the genomes of every clade in Fig. 1C. We found a total of 248 genes on chromosome I and 153 on chromosome II that showed signs of positive selection (dN/dS>1) (Fig. 6C). We reasoned that those genes could possibly serve as genetic variations associated with pathogen emergence (GVPEs). This recently coined term encompasses environmental preadaptations such as virulence adaptive polymorphisms and host-associated pathoadaptive mutations. Interestingly, our analyses did not retrieve a single gene that had a dN/dS>1 when compared to every other clade. Nonetheless, we identified a total of eight genes in PCG (four from each chromosome) that exhibit positive selection against more than 60 clades (Fig. 6C). Six of these genes (VC_0714, VC_1587, VC_A0224, VC_A0233, VC_A0609 and VC_A0973) had no match in the databases and therefore no function could be inferred, however, five of them encode transmembrane domains or signal peptides indicating a possible functional relationship with environmental sensing or interaction, either as secreted molecules, as receptors, channels, or membrane regulators. The remaining two genes encode a 2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate synthase (VC_1974), which is a hydrolytic enzyme involved in vitamin K synthesis, as well as, a NifB/NifX family molybdenum-iron cluster-binding protein (VC_1323), a cofactor of the enzyme dinitrogenase involved in the reduction of nitrogen to ammonium (Fig. 6C). There is no clear pattern regarding the functions associated with these genes, however, it is remarkable that, out of the over 200 clades analyzed, the PCG strongly exhibit signatures of selection for these 8. We
hypothesized that many of them could have arisen as GVPEs such as pathoadaptive mutations in the human host after the emergence of toxigenic strains from PCG.

CONCLUSIONS

In this study, we took a systematic approach towards unraveling the evolutionary history of the enigmatic Pandemic cholera group (PCG), which, to date, remains the only clade within the *V. cholerae* species capable of causing the severe diarrheal disease cholera. Phylogenomic analyses of the species reveal a cluster-based evolution of *V. cholerae* with several hierarchical levels leading to the emergence of the PCG, which is contained within a separate lineage (PCL) embedded within cluster 24. Our investigations on the abundance and distribution of the virulence-associated genes and MGEs suggest the existence of several layers in the evolutionary dynamics and emergence of *V. cholerae* (Fig. 7).

We determined that even the hypervariable region of the LPS is found in a large number of clusters. The high prevalence of the LPS cluster may be due to the different exchange mechanism of this island compared to the acquired MGEs. The LPS locus exhibit characteristics of a “replacement island”, as it is transferred as a unit by homologous recombination and not in small units like the other MGEs (Fig. 7A). On the other hand, complete islands are composed of several modules generally bordered by IS elements that contain small groups of genes resembling “additive islands” (Fig. 7B). Furthermore, the small gene groups within these additive islands also demonstrate allelic exchange within them and some can be fully conserved in synteny and similarity, whereas a very limited number of them are exclusive to PCG (Fig. 7C). Our results suggest that in order to analyze the evolutionary history of these elements, rather than investigating presence or absence of the island, it is necessary to examine the specific modules and individual alleles of some genes within them (Fig. 7). The evolution of the PCL shows that we can only trace the phylogeny of the LPS cluster, the only island that is exchanged by homologous recombination and which appears to have slower
replacement dynamics than the MGEs. However, both the example of the PCL and the phylogeny of the different modules in the whole species suggest that the rest of the islands have a rapid turnover and remain associated with each clade for a short time. Overall, this fractal-like degree of molecular sophistication explains the absence of a complete set of PCG-like elements in any other *V. cholerae* clade.

Genetic variations associated with pathogen emergence (GVPEs) in the form of environmental preadaptations or intra-host pathoadaptive mutations are becoming increasingly appreciated as major agents in the transition that allow bacteria to emerge as human pathogens. We examined the core genome microdiversity of *V. cholerae* and determined the large extent of allelic variations of genes that uniquely exhibit strong signatures of positive selection in PCG. This finding strongly suggests that, besides the intricate mosaic associated with MGEs modularity, unique allelic variations in PCG could help explain this confined distribution (Fig. 7D). Future work will determine whether these GVPEs stem from host adaptations or were present during the early evolution of the PCG. Overall, our investigations provide unique insights on the paths that led to the emergence of a major human pathogen and can extrapolated to other bacterial threats.

**FIGURE LEGENDS**

*Fig. 1. Phylogenomic and population structure of Vibrio cholerae.* **A.** Map of Florida indicating the locations where the environmental *V. cholerae* strains were isolated (Shepard Park and Feller’s House). **B.** Whole-genome SNP-based phylogenetic tree showing the relationships among the *V. cholerae* genomes obtained in this study and a representative strain from the Pandemic Cholera Group (*V. cholerae* O1 biovar El Tor N16961; red branch) **C.** Phylogenomic tree of dereplicated *V. cholerae* strains available in public databases and strains from this study. The tree was reconstructed from single nucleotide polymorphisms of the core genome. Branches of members belonging to the four main clusters are colored in green (C5), violet (C11), blue (C16) and red (C24). The
colored circles represent the source of isolation of the environmental strains collected in this study. Grey circles located at the branch nodes represent the bootstrap values.

Fig. 2. Distribution and evolutionary dynamics of the O1 Lipopolysaccharide (LPS) cluster in the *V. cholerae* species. **A.** The area outlined in gray represents the number of homologous proteins found in the dereplicated *V. cholerae* genomes. The white circles with dotted lines indicate the median protein identity. **B.** Genomic comparison of PCG-like LPS clusters identified in *V. cholerae* environmental genomes. **C.** Schematic representation of the variable zone of the LPS containing the *rfbT* gene and the IS elements inserted in the PCG allele. IS elements are highlighted in yellow.

Fig. 3. Integron-like dynamics govern genetic variation of CTXΦ phage. **A.** Distribution and structure of the CTXΦ phage in the *V. cholerae* species. The area outlined in gray represents the number of homologous proteins found in the dereplicated *V. cholerae* genomes. The white circles with dotted lines indicate the median protein identity. Red rectangles on the genes indicate their exclusive presence in the PCL. The black arrows indicate the genomic position where the cassettes have been inserted. **B.** Genomic comparison of similar CTXΦ phage found in dereplicated *V. cholerae* genomes. **C.** Schematic representation CTXΦ phage found in *V. cholerae* 3528-08 both the numbers and the rectangles mark the different clusters acquired. Clusters of genes with a known function follow the same color pattern. IS elements are highlighted in yellow.

Fig. 4. Modularity and allelic variations shape the evolutionary dynamics of Vibrio pathogenicity islands 1 and 2. Structures of **(A)** VPI-1 and **(F)** VPI-2. The area outlined in gray represents the number of homologous proteins found in the dereplicated *V. cholerae* genomes. The white circles with dotted lines indicate the median protein identity. **B.** Genomic comparison of homologous VPI-1 in dereplicated *V. cholerae*
genomes denote a reduced number of allelic variations in some PCG genes. Sequence alignment of the protein sequences of (C) ToxT (D) TcpA and (E) TcpF from PCG and IRLE0081 (C17). G. Genomic comparison of VPI-2 variants encoded in dereplicated V. cholerae genomes highlight pervasive modularity of the island. Clusters of genes with a known function follow the same color pattern. IS elements are highlighted in yellow.

Fig. 5. Distribution and evolutionary dynamics of the Vibrio seventh pandemic islands I and II. Structures of (A) VSP-I and (C) VSP-II. The gray area represents the number of homologous proteins found in the dereplicated V. cholerae genomes. The white circles with dotted lines indicate the median protein identity. Red rectangles on the genes indicate their exclusive presence in the PCL. Pink triangles indicate the presence of tRNA-direct repeat genomic islands. B. Genomic comparison of homologous VSP-I across V. cholerae genomes. D. Genomic comparison of VSP-II variants encoded in dereplicated V. cholerae genomes highlight pervasive modularity of the island. IS elements are highlighted in yellow.

Fig. 6. Evolutionary trajectories and genome dynamics of V. cholerae PCG. A. Pangeneome analysis. The Venn diagram shows the numbers of unique and shared gene families between the pandemic cholera group (purple), pandemic cholera lineage (blue), Cluster 24 (light blue) and the rest of clades (greyQ) B. PCL evolution. The tree designates the evolutionary relationships of the different clades within the PCL highlighting the loss and gain of genes through the lineage. Black numbers indicate an overall gain and grey an overall loss. The colored circles represent the specific element as well as the percentage of completeness in each of the clades in reference to PCG. LPS, O1 lipopolysaccharide cluster; VPI-1, Vibrio pathogenicity island-1; VPI-2, Vibrio pathogenicity island-2; VSP-1, Vibrio seventh pandemic island-1; VSP-2, Vibrio seventh pandemic island-2. C. Microdiversity analysis. PCG genes under positive selection
compared to the rest of the clades. X-axis indicates the number of clades in which $dN/dS>1$ and the Y-axis shows the average of the $dN/dS$ value.

**Fig. 7. Hierarchical model of the emergence of the Pandemic cholera group in *V. cholerae***. The evolutionary history of PCG suggests that the emergence of this group has not been a linear process. Our analyses indicate that this process has unfolded at four levels: A) The first one involving only the LPS where the island is replaced as a unit by homologous recombination (“replacement island”). This replacement is slow and allows tracing its phylogeny. B) The second level is related to the MGEs, which are islands of the additive type composed of several modules bordered by IS elements that have a very rapid exchange dynamics. In addition, these modules have the potential to be inserted into other regions of the genome. This rapid exchange precludes from associating these modules to specific clades. This is exemplified by the MGEs such as VPI-2 and VSP-II. C) Furthermore, a third level in the process of PCG emergence is associated with allelic variations within these modules, such as those we identified in the genes encoding TcpF or ToxT in VPI-1. D) Finally, the fourth level is composed of genetic variations associated with pathogen emergence (GVPEs), which are mutations in core genes associated with the emergence of pathogenic strains (e.g. VC1323 or VC0714) and exhibit strong positive selection in the PCG.

**SUPPLEMENTARY MATERIAL**

**Table S1.** List of *Vibrio cholerae* collected in this study, together with genomic features and metadata

**Table S2.** List of dereplicated (ANI>99%) *Vibrio cholerae* reference strains from the NCBI

**Table S3.** List of dereplicated (ANI>98%) *Vibrio cholerae* reference including IRL strains

**Table S4.** Shared family gene clusters in the pangenomes of the three hierarchical levels
Table S5. Summary of the gene gain/loss of the node 3 from the pandemic cholera lineage.

Figure S1. Dendrogram representing the relatedness of V. cholerae isolates. The dendrogram is based on PCR fingerprints generated by VCR-PCR. Groups of similarity were established using the UPGMA method in BioNumerics v8.0. The red circles indicate clonal isolates that were selected for whole genome sequencing.

Figure S2. Pairwise comparison among the V. cholerae genomes. The pairwise comparison was obtained using the average nucleotide identity (ANI) from strains isolated in this study and the Pandemic Cholera Group (V. cholerae O1 biovar El Tor).

Figure S3. Genomic comparison of CTXΦ phage. Comparison between the CTXΦ phage genomic island encoded by Pandemic Cholera Group strains (V. cholerae O1 biovar El Tor N16961) and Vibrio mimicus.

MATERIAL AND METHODS

Sampling sites. We collected samples at two environmentally distinctive locations along the IRL (Easter Florida, USA) in three sampling events. For each sampling campaign we obtained samples from several fractions: three associated with the water bodies (Water filtered through 20μm, 5μm, and 0.22μm) and one from the sediment. One of the sampling locations, Fellers House Field Station (N28˚54’25.315”; W80˚49’15.017”), is located within the federally-protected Canaveral National Seashore. The second sampling site, Shepard Park, is located in the Port St. Lucie area (N27˚11’48.864”; W80˚15’33.172”), due to urbanization and agricultural expansion, experiences nutrient over-enrichment leading to excessive macroalgal bloom (Fig. 1A)43.

Isolation of V. cholerae. Water. Isolation of V. cholerae from water samples collected at the aforementioned sampling locations was performed using a modified protocol from Huq et al71. Briefly, water samples (~3L) were collected in triplicate from each site and stored cold until arrival at the laboratory. To investigate the prevalence of V. cholerae in
various fractions of the aquatic column and to separate particle-associated and free-
living bacteria, 500 mL of collected water was filtered successively through 20µm, 5µm
and 0.2µm filters (Sterlitech) using a vacuum filtration system. The respective membrane
filters harboring potential *vibrio* spp. were then suspended in 25ml phosphate buffered
saline (PBS) and vortexed vigorously to release the contents. Each resuspended sample
was then inoculated (1:10 v/v) in alkaline peptone water (APW) and incubated at 37°C
for 12-14 hrs. This was repeated for each replicate. Enriched cultures were subjected to
growth on a series of selective agar specific for the isolation of *V. cholerae*. First,
enriched cultures were serially diluted 10-fold and plated on CHROMagar Vibrio (CaV;
CHROMagar), a chromagenic media on which *V. cholerae* forms turquoise blue colonies.
Secondly, turquoise blue colonies from CaV were transferred onto Thiosulfate Citrate
Bile Salts Sucrose (TCBS; Sigma) agar plates. Colonies that appeared turquoise blue
on CaV followed by yellow on TCBS were presumed to be *V. cholerae*. Sediment.
Isolation of *V. cholerae* was done using previous methods. Briefly, sediment samples
were collected in triplicates from each sampling site using universal corer. 25g of the
sediment samples from each replicate was weighed and suspended in 25 ml of PBS
(Gibco). Samples were homogenized, enriched, and studied for the presence of *V.
cholerae* as described above. *Cyanobacteria*. Cyanobacteria floccules were pelleted,
supernatant removed and fresh APW added to homogenize the sample. Samples were
enriched and studied for the presence of *V. cholerae* as described above.

**Strains and growth conditions.** *V. cholerae* strains were routinely grown on Luria
Bertani (LB) agar at 37 °C for approximately 16hrs, unless otherwise stated. For liquid
cultures, isolated colonies selected from agar plates were grown aerobically at 37°C in
LB broth for ~16hrs.

**Clonal analysis of *V. cholerae* isolates.** To confirm isolates of *V. cholerae*, presence
of the species-specific gene *ompW* was determined by PCR. Genomic DNA from the
isolates was extracted using the Gentra Puregene Yeast/Bact. Kit (Qiagen) and used as template for PCRs. To identify the clonal populations of *V. cholerae* in the confirmed dataset, DNA fingerprinting patterns were examined using a combination of *Vibrio cholerae* Repeats-PCR (VCR-PCR) and Enterobacterial Repetitive Intergenic Consensus Sequence (ERIC)-PCR as described by Teh *et al.* Genomic DNA was used to perform VCR- and ERIC-PCR and resulting PCR products were electrophoresed and imaged using UVP ChemStudio (AnalytikJena) to observe fingerprinting patterns. Cluster analysis of DNA patterns was performed using BioNumerics v8.0 (Applied Maths, Inc.) and dendrogram generated based on the Unweighted Pair Group Method with the Arithmetic Mean (UPGMA) method. Based on resulting DNA fingerprinting patterns, a total of 35 independent clones were identified and selected for whole genome sequencing (WGS). Genomic DNA of each clone was submitted to Microbial Genome Sequencing Center (MiGS) for WGS using the Illumina NextSeq 2000 platform.

**Genome assembly, gene prediction, and annotation.** Reads were trimmed using Trimmomatic v0.36 and assembled de novo with SPAdes version 3.11.1. Coding DNA sequences (CDS) from the assembled contigs were predicted using Prodigal version 2.6.3 using `-a output.proteins -d output.genes -c -p meta` parameters. Then, tRNAs were obtained using tRNAscan-SE version 1.4 together with ssu-align version 0.1.1, and rRNA genes using meta-rna. Predicted proteins were compared against the National Center for Biotechnology Information nonredundant database (NCBI nr) using DIAMOND. In addition, for functional annotation we used HMMscan version 3.1b2 for the comparison against COG v2003 (update 2014) and TIGFRAM v15.0 (September 2014) databases.

**Recovery of *V. cholerae* genomes.** A total of 1,127 *V. cholerae* genomes were downloaded from RefSeq (accessed July 2021) and subjected to a dereplication step with drep at 99 % nucleotide identity. Briefly, we force dRep to sort and select as seeds
for the clustering genomes with better score. For instance, chromosomes in a single contig, followed by higher N50 and L50 values, and lastly by the completeness and contamination values calculated with checkm\textsuperscript{84}. This resulted in 227 dereplicated reference \textit{V. cholerae} genomes.

**Phylogeny of \textit{V. cholerae} dereplicated genomes.** Single nucleotide polymorphisms (SNPs) among \textit{V. cholerae} dereplicated genomes and the PCG reference genome \textit{Vibrio cholerae} El Tor N16961 (RefSeq accession number GCF\_900205735.1) were calculated with PhaME\textsuperscript{85} with default options, resulting in a core genome length of 2,468,623 bp and 458,798 polymorphic positions considered. Then, a maximum likelihood phylogenetic tree of SNPs was constructed using iqtree v 1.6.12\textsuperscript{86} using the ultrafast bootstrap approach (5000 replicates) and the best fitted model GTR+F+R4.

**Gene cluster search.** To determine the distribution and degree of prevalence of proteins within the studied \textit{V. cholerae} genomes, we extracted the LPS cluster, CTXΦ and pathogenicity islands from the reference PCG strain based on the manual annotation found in the NCBI Genbank. Proteins were isolated from the clusters and analyzed using Prodigal\textsuperscript{76}, then used as a reference in a BLASTP search against the proteins from the genomes\textsuperscript{87,88}. We kept all proteins that matched against the pathogenic protein database with at least 70% identity and the alignment covered at least 70% of the both proteins.

**Pangenome analysis.** To evaluate the pangenome of \textit{V. cholerae}, in a first approach dereplicated sequences were grouped in clusters at 98% identity using cd-hit\textsuperscript{89}, resulting in 34 clusters, four of them (C5, C11, C16, and C24) comprising 188 out of 262 genomes. Then, pangenomes were calculated for each cluster individually using PPanGGOLiN\textsuperscript{90} with default parameters. A second pangenome step with the protein families of clusters C5, C11 y C16 was performed using PPanGGOLiN to simplify the comparison among the pathogenic reference genome el Tor, C24 and the new superpangenome C5-C11-
C16. Ortholog proteins among the three groups (el Tor, C24 and C5-C11-C16 were calculated using orthoMCL \(^{91}\) (> 30 % amino acid identity, > 70 % alignment). A Venn diagram showing proteins shared and not shared among groups was performed in R using the ggVennDiagram package (https://github.com/gaospecial/ggVennDiagram).

**Determination of Gain and Loss of genes in the pandemic cholera lineage.** In order to understand the flow of genes among pandemic *V. cholerae* and the closest genomes from cluster C24, those belonging to the pandemic cholera lineage, gene presence and absence was tested using the gain loss mapping engine GLOOME \(^{92}\). The expectations and probabilities of both gain and loss events were estimated using stochastic mapping. The phyletic pattern of presence/absence of orthologous genes was then classified using hmmscan \(^{90}\) against the Pfam database v35.0 \(^{93}\), using an evalue of < 1e\(^{-5}\) as a threshold for protein annotation.

**Microdiversity analysis.** To identify which genes in each genome were under positive selection, we calculated the nonsynonymous (d\(N\)) and synonymous (d\(S\)) substitutions between ortholog proteins found in both the studied *V. cholerae* genomes and the reference PCG strain, we used the orthologr package in R \(^{42}\), a pipeline that identifies orthologous protein pairs, performs codon alignment and finally computes the final d\(N/dS\) values. We considered a gene under positive selection if the d\(N/dS\) ratio > 1.

**ACKNOWLEDGEMENTS**

This work was supported by an NSF CAREER award (#2045671) to SAM. This work was also supported by grant “FLEX3GEN” PID2020-118052GB-I00 (cofounded with FEDER funds) from the Spanish Ministerio de Economía, Industria y Competitividad to MLP.

**COMPETING INTERESTS**

The authors declare no competing interests.
AUTHOR CONTRIBUTIONS

SAM conceived the study. MLP, AZS and JHM performed bioinformatic analyses. TAG collected samples, isolated *V. cholerae* and sequenced genomes. The manuscript was written by MLP and SAM. All authors read and approved the final version.

REFERENCES


35. O'Shea, Y. A. et al. The Vibrio seventh pandemic island-II is a 26.9 kb genomic island present in Vibrio cholerae El Tor and O139 serogroup isolates that shows homology to a 43.4 kb genomic island in V. vulnificus. *Microbiology* 150, 4053–4063 (2004).


63. Kirn, T. J. & Taylor, R. K. TcpF is a soluble colonization factor and protective antigen secreted by El Tor and classical O1 and O139 Vibrio cholerae serogroups. *Infection and immunity* 73, 4461–4470 (2005).


Functional Analysis of the 18 Vibrio cholerae N16961 Toxin-Antitoxin Systems

71. Huq, A. *et al.* Detection, Isolation, and Identification of Vibrio cholerae from the

72. Naha, A. *et al.* Development and Evaluation of a PCR Assay for Tracking the
Emergence and Dissemination of Haitian Variant ctxB in Vibrio cholerae O1 Strains

73. Zou, W. *et al.* Evaluation of Pulsed-Field Gel Electrophoresis Profiles for


75. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its
Applications to Single-Cell Sequencing. *Journal of Computational Biology* 19, 455–477
(2012).

76. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site

77. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of

78. Huang, Y., Gilna, P. & Li, W. Identification of ribosomal RNA genes in metagenomic


80. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Computational Biology* 7,
e1002195 (2011).

81. Tatusov, R. L. *et al.* The COG database: new developments in phylogenetic
classification of proteins from complete genomes. *Nucleic acids research* 29, 22–28

82. Haft, D. H. *et al.* TIGRFAMs: a protein family resource for the functional

accurate genomic comparisons that enables improved genome recovery from

84. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W.
CheckM: assessing the quality of microbial genomes recovered from isolates, single


Pandemic cholera group (PCG)

Source of isolation:
- Water. Particle-associated (≥5µm)
- Water. Free-living (≤0,2µm)
- Sediment
- Cyanobacteria

Florida

Tree scale: 0.1 µm
Bootstrap (%)

0.75 50

Pandemic cholera group (PCG)